DOCUMENT RESUME

ED 178 599 TM 009 806

AUTHOR Gillmore, Gerald M.
TITLE An Introduction to Generalizability Theory as a Contributor to Evaluation Research.
INSTITUTION Washington Univ., Seattle. Educational Assessment Center.
REPORT NO EAC-R-79-14
PUB DATE Mar 79
NOTE 56p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Analysis of Variance; Criterion Referenced Tests; Decision Making; Educational Assessment; Educational Research; *Evaluation; Item Sampling; Mastery Tests; Norm Referenced Tests; *Reliability; *Research Design; Sampling; *Test Reliability; Tests; Test Validity
IDENTIFIERS Classical Test Theory; *Generalizability Theory

ABSTRACT

It is argued in this paper that generalizability theory provides a uniquely useful framework for defining and quantifying the dependability of data for decision making. It does so by requiring careful specification of the conditions of measurement and the anticipated sources of variation in the results of the measurement procedure. A distinction is made between generalizability (G) studies and decision (D) studies. A G study is conducted to determine certain characteristics of the measuring device, while a D study is conducted in support of decision making. The decision-maker must specify (1) the object of measurement; (2) the universe of generalization; (3) the conditions to be sampled from each facet; and (4) the experimental design. Summary statistics for D studies may include: the expected observed score variance, the universe score variance, three types of error variance, and indices of dependability, including the generalizability coefficient. An application of the generalizability coefficient is described and discussed. Confidence intervals are discussed briefly. The testing of persons as individuals, rather than as class members, is discussed and compared for norm referenced and domain referenced tests and specifically for mastery tests. (CTM)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

EDUCATIONAL ASSESSMENT CENTER

UNIVERSITY OF WASHINGTON

March, 1979

# An Introduction to Generalizability Theory as a Contributor to Evaluation Research

79-14

Gerald M. Gillmore

ABSTRACT

The purpose of this report is to introduce the basic principles of generalizability theory and to illustrate their application to evaluation research. Following definition of G studies, D studies, and D study summary statistics, special attention is given to student instructional ratings as an example application. The report concludes with discussion of norm-referenced, domain-referenced, and mastery testing and of concerns relating to the validity of measures.

---

EAC Report Number: 79-14

EAC Study Number: 506

# An Introduction to Generalizability Theory as a Contributor to Evaluation Research[1]

Gerald M. Gillmore

Edwards, Guttentag, and Snapper (1975) have claimed that "Evaluations... exist (or perhaps only should exist) to facilitate intelligent decision-making" (p. 140). They, in fact, described the unique character of evaluation research in terms of its relationship to decision-making. "...if [evaluation research] does not improve the basis for decisions about the program and its competitors, then it loses its distinctive character as evaluation research and becomes simply research" (p. 140). In reality, the relationship between evaluative research results and real decisions is seldom intimate. This is partially due, no doubt, to the ever present political considerations which surround decisions (Weise, 1975). It is also due, or should be, to the twin uncertainties which haunt the measurements of social science and educational researchers; namely, uncertainty as to what is being measured and how well it is being measured. Where decisions are based on data, the quality of those data, vis a vis the specific decision they are intended to support, is a highly relevant concern. All other things equal, better data lead to better and more confident decisions.

Strictly speaking, the measurements obtained in conducting evaluative research, and most other types of research for that matter, are seldom the measurements desired. This reality has led, directly or indirectly, to the large measurement theory literature and recently to generalizability theory. Consider, for example, a typical classroom testing situation. A teacher might

wish to assess the ability of each student to, say, multiply fractions. To this end, a twenty-item test is administered which in the teacher's opinion represents the behavior sought. The result of this measurement procedure is a score for each student representing the number or percentage of the twenty items answered correctly. The teacher uses this score, and perhaps other scores and impressions, to judge the student's competence in multiplying fractions. Clearly, a score on a twenty-item test is not what the teacher really wants upon which to base decisions. A quantification of the student's ability to multiply fractions is desired; a student's ability to successfully answer a relatively small number of relevant problems at one distinct point in time is obtained. If, on a subsequent day, this same teacher were to administer a different twenty-item test covering the same subject, the resulting scores would undoubtedly show some change. As Segal's law states: "A man with one watch knows what time it is. A man with two watches is never sure" (Wallace, Wallechinsky, and Wallace, 1977, p. 480).

One way to view this example is to note that in essence the teacher has the results of a twenty-item test, administered on a specific occasion; but these results must be used as if, in making decisions about the student's ability to multiply fractions, they were based on all possible items covering this ability and on all acceptable occasions at which the test could be given.[2] The extent of errors from imperfect item selection, poor timing of the test, and other sources will determine how well specific results represent the actual behavior sought, which will, in turn, strongly influence the quality of the decisions made.

What is argued in this paper is that generalizability theory provides a uniquely useful framework for defining and quantifying the dependability

of data for decision-making. To anticipate later discussion, it does so by requiring careful specification of the conditions of the measurement and the sources of variation in the results of the measurement procedure. Furthermore, generalizability theory requires explicit specification of the actual decisions which the data will support and evaluates the measuring device with regard to these decisions. The burden is clearly placed on the decision-maker for specifying the object of the decision and the conditions which are represented by the data.

## Background

The theoretical foundations of generalizability theory can be found in Cronbach, Rajaratnam, and Gleser (1963) and Gleser, Cronbach, and Rajaratnam (1965). The most important presentation of the theory to date is the book by Cronbach, Gleser, Nanda, and Rajaratnam (1972) entitled, The Dependability of Behavioral Measurements. The monograph of Brennan (1977) is a significant addition to the above for aiding practitioners in the actual conduct of generalizability studies.

The roots of generalizability theory are varied and have been elaborated elsewhere (Cronbach, et al, 1972, pp. 5-14). Briefly, the theory owes its lineage to two major traditions: classical test theory (Gulliksen, 1950; Lord and Novick, 1968) and the analysis of variance (Fisher, 1925). However, although generalizability theory may owe its inception to these two parents, it enjoys a different relationship with each. Generalizability theory takes advantage of the methodological techniques associated with the analysis of variance, with particular interest in the design of experiments and in variance component estimation and relatively little interest in more traditional hypothesis testing applications. In regard to classical test theory, it is

this writer's opinion that generalizability theory supercedes this prior conception through its superior conceptualization and integration, through its stress on the decisions which the measurements support, and through its ability to take advantage of complex experimental designs. In fact, for all practical purposes classical test theory can be viewed as a special case within generalizability theory.

In classical test theory, the total observed score variance in a given study is partitioned into two components: true score variance, reflective of "true" differences among persons or collections of persons, and error variance, reflective of undifferentiated random error. The reliability of the measuring instrument is defined as the ratio of true score variance to total observed score variance or one minus the ratio of error variance to total observed score variance. No matter how a reliability study is conducted (e.g., test-retest, equivalent forms, split-halves) these two components emerge. Clearly, the definitions and meaning of the two terms can vary considerably from situation to situation and within varying methods of estimating reliability. In spite of this, or perhaps because of it, the concept of a true score often enjoys a mystical, reified quality as if it were something a person or other entity really possesses but the magnitude of which only God knows for certain.

Like true score, the concept of error in classical test theory is also assumed to be undifferentiated and unvariate, even though there is often explicit recognition that the errors can emanate from multiple sources. This undifferentiated view of error can lead to misperceptions of the quality of one's measures. Furthermore, a person who wishes to base decisions upon the data from a measuring device may be doing so within a context dissimilar to that in which the reliability was estimated. The true score definition

may be inappropriate and the sources of error might be different. As a result, the decision-maker can be easily fooled into accepting an estimate of reliability which is inappropriate for the given decision to be made. As an example, in many evaluation studies nationally standardized test results are used as a dependent variable. Frequently, the reliability and validity statistics published by the test producer are nearly irrelevant for judging the adequacy of the test for the purpose to which it is put. This is not so much a failure on the part of the test publisher but rather an indication that classical test theory, because of its imprecise and inflexible definition of true score and error is not well-suited to the needs of evaluation research.

As we shall see, use of the conceptual framework of generalizability theory minimizes these confusions and greatly enhances the amount of information derivable from a multifactor study. In generalizability theory, the variation in the results of a measurement procedure are partitioned into as many sources as the design of the study allows which is often greater than the two allowed by classical test theory. Which of these sources is appropriately considered replicable variation (true score variance in classical test theory terminology and universe score variation in generalizability theory terminology) and which is appropriately considered random error depends upon the purpose to which the data are to be put. In fact, these definitions can change as the same measurements serve different purposes, e.g., to discriminate among persons or to discriminate among classes of persons. As with situation ethics (Fletcher, 1966), generalizability theory recognizes that any source of variation is not inherently good, bad, or indifferent. It all depends on the context.

## Formulation

To understand precisely how generalizability theory can be useful to evaluation research, the reader must be acquainted with its basic elements which are presented below. The best explication of the theory remains Cronbach, et al (1972). Readers familiar with the book may find this paper useful as a review. Those for whom this is largely a first exposure may wish to consult Cronbach, et al for a more in-depth treatment of topics than can be accomplished here.

This presentation begins with discussion of generalizability or G studies followed by discussion of decision or D studies. The conceptual distinction between G studies and D studies was first introduced by Rajaratnam (1960), and is basic to an understanding and appreciation of generalizability theory. In essence, a G study is conducted to determine certain characteristics of one's measuring device and is most effectively performed in the instrument development stage. D studies, in contrast, are studies upon which actual decisions will be made. As generalizability analyses are done in support of evaluation research, D study results take on special importance since they are most directly related to the data-based decisions which are to be made. However, a well-conceived and conducted G study is crucial to meaningful D study conclusions. The evaluation researcher may or may not have the luxury of an existing appropriate G study. Either way the G study results have to be understood and transformed into specific D studies results. Thus, the formulation begins with an explanation of G studies.

### G Studies

The starting point of a G study is the definition of the universe of admissible observations. This concept refers to the factors or variables which

one chooses to take into consideration in the G study, and the range of acceptable conditions for each. Simply put, the universe of admissible observations is "...the observations encompassed by the possible combinations of conditions that the G study represents" (Cronbach, et al, 1972, p. 20). For example, one of the factors of the G study might be persons, and the universe of admissible observations for persons might be all sixth grade students within a certain large city school district. The actual G study conditions for the persons factor then should be a random sample from this universe. Other factors in the G study might be classes and test items. Each of these would also require careful definition of the universe of admissible observations.

The designs of G studies will virtually always contain repeated measures on persons; that is, persons are a factor (independent variable) in the design, and more than one measurement is taken on each person. Furthermore, in analysis of variance terminology, each person within the sample can be considered a level of the persons factor. G study designs

may be completely crossed, i.e., each level of a given factor is represented within all levels of the other factors, or may contain some factors which are nested within other factors. For example, in educational settings the persons factor is frequently nested within the classes factor. Each class contains a set of persons, but each person is found within only one class.

The goal of the typical G study is the estimation of variance components for all effects which are estimable from the design of the study. An in-depth treatment of techniques for calculating variance component estimates is beyond the scope of this chapter (see Brennan, 1977; pp. 22-27 for specific algorithms). However, the partitioning of the total variance from a study into components which are attributable to various sources is so fundamental to generalizability theory that some treatment will be given here.

Consider a relatively simple design in which each person within a sample responds to the items of a test. The design of this illustrative study is simply persons crossed with items, which can be denoted p x i. An observed score, denoted $X_{pi}$, can be defined as the score of person p on item i. If these items are scored dichotomously (right or wrong) then $X_{pi}$ can equal either one or zero for each unique person p and item i. The variance will be equal to .25 if item difficulty equals .5, or be less than .25 as the average departs from .5. For the p x i design, this total variance of the scores can be partitioned into that due to the item main effect, that due to the person main effect, and that due to the person by item interaction. Of course, we need not restrict ourselves to either dichotomously scored items or the p x i design. For example, if raters, r, were crossed with both persons and items, perhaps in an application in which essay examination questions were graded by several raters, variance component estimates could be obtained for

the main effects of raters, persons, and items, the three two-way interactions, pi, pr, and ir, and the three-way interaction, pir.

Nested G study designs result in the confounding of certain variance component estimates. To illustrate, consider a design in which items are nested within persons rather than being crossed with persons. In this design, the variance components for the items main effect and for the items by persons interaction cannot be separately estimated. Instead, a single variance component estimate is obtained which contains both of these effects. As a general rule, the variance component of the main effect for a nested factor is confounded with the variance component of the interaction of that factor with the factor in which it is confounded. In nested factors within designs of more than two factors, similar confounding exists between the variance components of two-way and three-way interactions, three-way and four-way interactions, etc.

In most experimental design texts (e.g., Myers, 1966), expected mean squares are listed as functions of specified variance component sums largely to facilitate selection of the proper error term with which to test hypotheses about certain effects. In generalizability theory, the variance components themselves take on extreme importance. Whereas in classical test theory the variance is partitioned into that attributable to only two sources, true score variance and error variance, in generalizability theory a G study yields variance component estimates for all separately estimable effects. Variance components should not be confused with either sums of squares or mean squares found in typical analysis of variance summary tables, although variance components share the additive properties of the former and are estimated from appropriately weighted combinations of the latter.

Formulae for variance component calculations differ depending upon whether

factors are considered random or fixed. However, in the conduct of a G study, all variance components should be estimated assuming a completely random model. As we shall see, decisions as to how certain factors should be considered can be deferred until the D study is defined without loss of accuracy and with gain of useful information. Also, G study variance component estimates should be viewed as being for average single observations, e.g., an average person responding to an average item. The actual number of conditions to be sampled for decision-making is defined by the D study.

D Studies

One could say that the general characteristics of the measuring device, be it a test, rating instrument, or whatever, are established in conduct of a G study. The D study, then, is conducted to assess the specific characteristics of the measuring device in a particular decision-making context. In most cases, the conduct of a D study will involve the collection of new data. A typical and efficient practice in evaluation research is for the G study to be essentially a pilot testing of the instrument with an appropriate random sample. These data may be then used to simulate D study results, but the actual D study, conducted in support of decision-making, might involve measurements on a much larger sample or on an entire population. However, real circumstances might dictate that the G and D studies involve the same data. In either case, the D study definition is dictated by how the data will be interpreted when they are used to make decisions. This definition can be viewed as the result of four decisions. The decision-maker must specify 1) the object of measurement, 2) the universe of generalization, 3) the number of conditions to be sampled from each facet, and 4) the experimental design. Each of these will be discussed in turn.

1. The object of measurement. The object of measurement for a D study is that G study factor about which decisions are to be made. Frequently, the object of measurement will be persons, but it may be some object other than persons as well. For example, consider a standard testing situation in which persons are crossed with items. If the test is being used to make judgments about persons, perhaps to assign grades, the object of measurement for the D study should be persons. If in a different design, say where persons are nested within classes and crossed with items, the object of measurement is appropriately classes if the purpose of the measurement is to make judgments about classes, and is persons if the purpose of the measurement is to make judgments about persons.

Considerations relating to the definition of the object of measurement are very important within evaluation research because the object of measurement is often other than persons. Kane and Brennan (1976) cited evaluations of Head Start and Follow Through as examples of evaluations for which the class was an appropriate object of measurement. In presenting findings from the evaluation of Follow Through, Kennedy (1978) pointed out that "...Several levels of data aggregation were considered as units of analysis: the child, the classroom, the school building, and the community (the site)" (p. 3). Research on student instructional ratings has included generalizability studies in which the class, the teacher and the course have been objects of measurement (see Kane, Gillmore and Crooks, 1976 and Gillmore, Kane and Naccarato, 1978).

In cases where decisions are to be made about objects of measurement other than persons, assessments of the quality of one's measuring device for making decisions about persons are likely to be misleading. Consider as an example, the use of a college entrance examination to make decisions about the instructional program of schools. High reliability estimates are reported for these tests when administered to high school juniors and seniors, e.g.,

median subscale reliability estimates of .84 and above (KR-20) are reported for the ACT Assessment (American College Testing Program, 1973, p. 103). These estimates indicate that the variance of the scores of persons is large relative to the variance of the person by item interaction. If one's goal is to make discriminating decisions among persons (e.g., for college entrance), this relatively large variance among persons is desirable. But, if one wished to make discriminations among schools based on the test results of persons within schools, the above reliability estimates are no longer of interest. For this latter application, variance among persons (within schools) is not desired but becomes an additional source of error. For optimal discrimination one would wish all persons within each school to obtain nearly the same score. Thus, the same standardized tests which are successful in discriminating among persons will often fail to discriminate among larger units like schools because the differences among persons within schools tend to swamp any differences among schools. The reader is referred to Cardinet, Tourneur, and Allal (1976) for a more extensive discussion of questions which various choices of objects of measurement can address within a study.

2. Universe of generalization. Cronbach, et al (1972) used the term "facet" to denote what has been termed "factor" here, and used "conditions" to denote the levels of the facet. Their terminology is adopted in this paper for those factors which are included in the D study but are not the object of measurement. For example, if the factors of a study are persons and items and the object of measurement for the D study is persons, then the items factor would be a facet of the D study.

In determining the universe of generalization, at issue is the set of conditions to which the decision-maker wishes to generalize the resulting measures. In defining this set of conditions, each facet can be considered

as a random sample from an essentially infinite set of possible admissible conditions. For example, a random sample of sites might be selected for trial implementation of a compensatory program with sites being one of the facets of the D study. However, the entire essentially infinite set of possible sites at which the program could be implemented might be considered as within the universe of generalization. For a facet so considered, generalization of results is considered to be to the entire set of admissible conditions; for the above example, to the entire set of possible sites. Consistent with conventions in the analysis of variance literature, facets considered this way are termed random facets.

Second, the conditions of a facet can be taken as exhausting the entire set of admissible conditions for that facet, and generalization beyond the set is not of interest. For the above example, only the sites used in D study might be defined as within the universe of generalization for the sites facet. Facets considered in this way are termed fixed. Finally, the decision-maker may wish to generalize the results to a set of conditions for a facet which is larger than that sampled in the D study, but which is clearly not infinite. For example, a random sample of twenty states might be chosen in which to try out a state-wide program, and decision-makers might wish to generalize the results to all fifty states. In such a case, generalization is intended to be to a finite universe. This latter case is unusual within generalizability theory and in this paper attention will be restricted to fixed and random universes of generalization. Procedures for conducting D studies with finite universes of generalization are presented by Brennan (1977, pp. 49-57), and the interested reader is referred there.

The universe of generalization for a given D study is highly related to the universe of admissible observations, defined in the G study. The two can be identical or the former can be a subset of the latter. For example,

certain factors of the G study could be collapsed and not defined as facets for the D study, perhaps because the associated variance components were negligible. Importantly, the universe of generalization cannot be larger in any respect than the universe of admissible observations--it cannot include facets not present as factors in the G study and it cannot define the universe of generalization for a facet which is not completely contained in the corresponding universe of admissible observations of the G study.

Perhaps it would be useful to further illustrate the distinction between fixed and random universes with an additional example. Consider the design in which persons are nested within classes and crossed with items. Following the notation of Brennan (1977) this design can be denoted (p:c) x i. The colon designates a nested effect and should be read "nested within." As will be discussed subsequently, this is the design under which student instructional ratings are typically collected. Assuming that the object of measurement is classes for a D study employing this design, the facets of the study would be persons and items. Consider persons first. The persons facet should be considered as random if the decision-maker wishes to generalize the results of a class to the essentially infinite set of persons who could be considered as admissible members of the class. Those persons actually in the class would be considered a random sample from this infinite set. If, however, the decision-maker does not wish to generalize the results beyond those specific persons within the class, persons should be treated as a fixed facet.

For purposes of illustration, let us say that this particular design is being used to evaluate the teachers of the classes by means of standardized test results. Treating students as a random facet implies that what is desired is a measure of student learning which goes beyond the learning of just those particular students within the class. What is of interest is the amount of learning which would be evidenced in classes composed of any admissible students. Undoubtedly, in most decision-making contexts this is exactly what is desired.

Treating students as fixed implies that the measure desired is the test results of the specific students within the class and generalization to some larger set is not desired.

The same choice arises for items, again dictated by the particular needs of the decision-maker. If the items facet is considered as fixed, the decision-maker is assenting to a focus on the results of those specific items and the universe of generalization contains only those items. Treating the items facet as random expands the universe of generalization to include the entire item domain from which the particular items can be considered a random sample. In the former case, the decision-maker is, in essence, indicating a satisfaction with the particular items used as completely specifying the desired behavior. In the latter case, the items are viewed as only a partial sample of the desired behavior.[3]

3. The number of conditions of each facet. In the G study, a certain number of conditions or levels of each factor are sampled in order to establish estimates for the variance components for each effect of the design. It is important in conducting a G study to choose an adequate number of conditions to obtain stable estimates (see Cronbach, et al, 1972, pp. 49-50), although no very good guidelines for making this choice are available. However, the number of conditions sampled for each facet in the D study is not restricted to the number sampled in the G study, but rather, must be specified by the decision-maker. (As we shall see later, an investigator can systematically vary the number of conditions to be sampled in the D study in order to forecast the resulting dependability.) As an example, in the persons by items design, a G study may have been conducted with fifty items and, thus, the number of levels for the item factor was fifty. But, the decision-maker may wish to, or may have to, base judgments on a test of only twenty items. Hence, the number of conditions for the item facet in the D study would be twenty and not fifty.

In this chapter, decisions made concerning the object of measurement are considered to be based on averages over all of the D study conditions of the facets. To illustrate, in the persons by items design, decisions about persons are based on a score for each person which is determined by averaging over all items. In the design in which persons are nested within classes and crossed with items, a decision is made about a class based on the average response of all persons within the class to all items. Results would be equivalent if the choice were to base decisions on sums rather than averages, as long as one is consistent throughout.

Recall that G study variance components are based on one average level for each effect. For the D study, the variance component associated with an effect is divided by the number of conditions of the effect over which a score is averaged within the object of measurement. In other words, the D study variance component for a given effect is equal to the corresponding G study variance component divided by the number of D study conditions for that effect within the object of measurement. G and D study variance components are equal for effects which only contain the object of measurement.

Table 1 contains the G study variance components and comparable D study variance components for the (p:c) x i design, assuming classes as the object of measurement and a balanced design. The number of items in the D study is denoted $n_i'$, and the number of persons within each class is denoted $n_p'$. The primes are included to denote the number of D study conditions.

Table 1 about here

To highlight a few of the entries in the table, one can notice that the variance component for the class main effect is not altered from the G study to the D study, since it contains only the object of measurement. The variance component for the classes by item interaction is divided by the number of items in the D study. The persons by items within classes effect (pi:c) is divided by both the number of items and the number of persons since this effect contains both of these facets, and it is the product of the number of conditions which is sampled within the object of measurement.

4. The experimental design. Most of what is to be said within this section has been anticipated earlier. Frequently, the experimental design for the D study will be identical to that of the G study. There are two exceptions which immediately come to mind. First, a factor which was

crossed with other factors in the G study may become a nested factor in the D study. A good example of this is a task requiring judgments of raters, such as in grading the written essays of persons. In the G study, all raters might rate all items on all essays, thus allowing independent estimates of the variance comporents due to raters, persons, and the raters by persons interaction. In the D study, essays might be nested within raters; thus, rater 1 might grade the first twenty essays, rater 2 the next twenty, etc.

There are two reasons why this strategy might be adopted. First, practical circumstances such as limited resources may require nesting. There may be too little money to hire more than one rater to do every essay. Secondly, the G study results may suggest that the change can be made with little loss in the resulting dependability of the data for decision-making. It is important to note that crossed design results can be collapsed into nested design results simply by summing appropriate variance component estimates. For the above example, the variance component estimate for the persons within raters effect is equal to the sum of the estimates for the rater effect and for the rater by person interaction.

The second way a G study design could differ from a D study design is by not including factors of the former in the design of the latter. For example, the age of the rater might be a factor of the G study, but in the D study all raters are chosen to be of the same age or rater age is left to vary without any attention given it. Again this change of design might be based on resource limitations. It might also be based on G study results. All variance components associated with the rater-age factor could be of negligible magnitude and, thus, be of little consequence in determining the outcomes of the measurements.

The reader should note that in both of the examples in which the D study design differed from the G study design, the former was a subset of the latter. To repeat, a D study cannot treat a facet or the object of measurement as crossed when it was nested within the G study. Nor can the D study include facets which were not factors in the G study. The reason in both cases is simple: all of the variance component estimates necessary for D study inferences would not be available from the G study.

D Study Summary Statistics

The presentation of the basic formulation of generalizability theory has thus far discussed G studies and stressed the importance of variance component estimates derived from them. Furthermore, four major choices facing the decision-maker in defining a D study were outlined. The decision-maker must decide upon the object of measurement, the universe of generalization, the number of conditions of each facet to be sampled, and, finally, the design under which D study data will be collected. When these decisions are made, and a G study has been conducted, D study variance components are directly estimable. As Brennan (1977) stated, "D study variance components are useful in and of themselves, because they provide a direct indication of the relative magnitude, in the D study, of each of the independently estimable components of score variance" (p. 31). However, certain summary statistics can be defined, all based on D study variance component estimates, which can be useful in assessing the measuring device. The following statistics will be discussed: the expected observed score variance, the universe score variance, three types of error variance, and indices of dependability, including the generalizability coefficient. The notation of Brennan (1977) with slight modification is used in defining these statistics.

1. The expected observed score variance, $\hat{\varepsilon}_o^2(X)$, is the expected value of observed squared deviation scores, with the expected value being taken over design replications. For the (p:c) x i design, with persons as the object of measurement, $\hat{\varepsilon}_o^2(X)$ is the expected value of the squared differences between each person's score, averaged over the facets of the study, and the mean of the scores of all persons within classes. Similarly, with classes as the object of measurement, it is the expected value of the squared differences between the mean score of each class and the mean score over all classes.

Operationally, the D study expected observed score variance can be estimated by summing all of the D study variance component estimates which contain, as one of their indices, the object of measurement index. In the (p:c) x i design (see Table 1), with classes as the object of measurement and with $n_i'$ items and $n_p'$ persons sampled in the D study, the expected observed score variance is estimated as follows:

Insert Equation 1 about here

Note that although the variance component for the main effect of items is estimable in the G study, and, hence, in the D study, it is not part of the expected observed score variance. This is because $\hat{\sigma}^2(i)$ does not contain the object of measurement index, c. All persons respond to the same items, and differences among the scores of classes are based on averages over those items. Hence, differences among item difficulty will have no effect on differences among observed scores of classes.

2. The universe score variance, $\sigma^2(\tau)$, is that portion of the expected observed score variance which would remain if samples of infinite size were taken for each random facet of the D study. It is analogous to the true score

variance of classical test theory. Variance components which form the universe score variance are those which contain the object of measurement index only, or contain the object of measurement and one or more fixed facet indices. No variance component which contains a random facet index can be part of the universe score variance, because, theoretically, this is not replicable variance in the universe of generalization.

An easy way to determine the universe score variance is to take the limit of the expected observed score variance as all sample sizes for random facets go to infinity. The D study variance components containing random facet indices then go to zero, and the non-zero terms represent the universe score variance. In the formula for the estimated observed score variance presented in equation 1, above, if both items and persons are considered as random facets, then the universe score variance would be estimated by $\hat{\sigma}^2(c)$ alone. If the items facet is considered random, but persons fixed, it is estimated by $\hat{\sigma}^2(c) + \hat{\sigma}^2(p{:}c)/n_p'$. If the persons facet is considered random, but items fixed, it is estimated by $\hat{\sigma}^2(c) + \hat{\sigma}^2(ci)/n_i'$. Finally, if both items and persons are considered as fixed facets, then there is no independently estimable error variance and, hence, the universe score variance is equivalent to the observed score variance. In this trivial case, there is no generalization beyond the obtained observed score.

3. The error variances, $\sigma^2(\delta)$, $\sigma^2(\Delta)$, or $\sigma^2(\varepsilon)$, all relate to average squared differences between observed and universe scores. The appropriate error variance to use in a given situation depends upon whether decisions will be based on the observed deviations from the sample mean, directly on the observed scores, or on a regression estimate, respectively. Each of these error variances will be discussed in turn.

a. The error variance for making comparative judgments, $\sigma^2(\delta)$, is appropriate when decisions are made on a relative rather than on an absolute basis and is based on the discrepancies between observed and universe deviation scores. It is that portion of the expected observed score variance which is not part of the universe score variance. The expected observed score variance, then, is the sum of the universe score variance and the error variance for making comparative judgments.

Operationally, the error variance for making comparative judgments is the sum of those D study variance components which contain at least one index of a random facet and the index for the object of measurement. Again using the (p:c) x i design with classes as the object of measurement as an example, the equation for estimating this error, with both persons and items considered random facets is as follows:

Insert Equation 2 here

If the persons facet is treated as fixed, then the variance component estimate associated with the persons within classes becomes a source of universe score variance rather than error variance, and the error variance is estimated as follows:

Insert Equation 3 here

Finally, if the items facet is treated as fixed, and persons facet is treated as random, then the formula is:

Insert Equation 4 here

b. The average error variance within the object of measurement, $\sigma^2(\Delta)$, is the variance of the differences between the observed scores and the universe scores. It is composed of all those variance components which contain one or more random facet indices. The average error variance within the object of measurement will contain all of the variance components of the

error variance for making comparative judgments plus those variance components which contain only random facets. In the (p:c) x i design from which equations 1 through 4 were formed, the following estimate for this error variance would hold, assuming both facets are random:

Insert Equation 5 here

Note that $\hat{\sigma}^2(i)/n_i'$ is contained in the above equation whereas it was absent in equation 2. In cases where the items facet is treated as fixed, $\hat{\sigma}^2(\delta)$ and $\hat{\sigma}^2(\Delta)$ are equivalent for this particular design. We shall see later, however, that for applications in which $\hat{\sigma}^2(\Delta)$ is the appropriate error term, items will seldom be viewed as a fixed facet.

The average error variance within the object of measurement is applicable when decisions about the object of measurement are to be made on an absolute basis rather than on a relative basis. If, for example, test results are to be interpreted relative to the domain from which the items were sampled rather than relative to the sources of other examinees, $\sigma^2(\Delta)$ is the appropriate error variance. This topic will be discussed in more detail in the subsequent section on the testing of persons.

c. The error variance of estimate from the linear regression of universe scores on observed scores, $\sigma^2(\varepsilon)$, is the variance of the differences between estimated universe scores and actual universe scores. This error variance estimate is appropriate when linear regression is used to take group membership variables along with the observed scores into account in estimating the observed scores. The equation for calculating $\sigma^2(\varepsilon)$ is as follows:

Insert Equation 6 here

Before using this estimation technique the reader is cautioned to consult Cronbach, et al (1972) for a more in-depth treatment. It is included here

mainly for completion and further discussion is beyond the intended scope of this chapter.

4. Indices of dependability are essentially squared correlations between universe and observed scores. The precise definition varies depending upon universe and observed score definitions. In general, dependability is operationally defined as the ratio of universe score variance to the sum of universe score variance and error variance. When decisions are made on a comparative basis and the appropriate error is $\delta$, the denominator for the index of dependability is the expected observed score variance. In this case, the index is termed the generalizability coefficient and is denoted $\mathcal{E}_\rho^2$. When decisions are made on an absolute basis, and the appropriate error is $\Delta$, the generalizability coefficient is likely to be an overestimate of dependability. The dependability index which is appropriate when decisions are made on an absolute basis will be discussed in a subsequent section on the testing of persons. A third index of dependability could be defined for cases in which the observed score is determined by use of linear regression techniques. However, such a definition is beyond the scope of this chapter.

Immediate attention will be given to the generalizability coefficient. This will be followed by an application, dealing with the evaluation of college-level instruction, in which generalizability coefficients are computed and interpreted. The subsequent section will take up the testing of persons and, within that context, discussion of indices of dependability will be extended.

The generalizability coefficient, $\mathcal{E}_\rho^2$, is analogous to the reliability

coefficient within classical test theory. In classical test theory, reliability is defined as the ratio of true score variance to observed score variance. In generalizability theory, generalizability is defined as the ratio of universe score variance to expected observed score variance. "It is completely comparable to the traditional reliability coefficient except that full attention has to be given to the universe definition and to the design of the D study" (Cronbach, et al, 1972, p. 97).

The generalizability coefficient is an intraclass correlation which can be interpreted as an average correlation, over the objects of measurements among successive replications of the D study design. Each of these replications would use the same objects of measurement (e.g., the same persons), the same conditions for fixed facets, and a new, same-sized random sample of conditions for random facets. The scores theoretically correlated are deviations from the overall mean for each replication. The generalizability coefficient can also be viewed as an approximation of the squared correlation between the observed deviation scores and the universe deviation scores.

For the (p:c) x i design with classes as the object of measurement which we have been following as an example, there are four distinct generalizability coefficients possible, depending upon whether the items facet is treated as fixed or random and whether the persons facet is treated as fixed or random. When both items and persons are treated as fixed facets, however, the generalizability coefficient is trivially 1.0 since there is no independent estimate of error variance and, thus, the estimate for universe score variance is the same as the estimate for the expected observed score variance. The formulae corresponding to the remaining possibilities follow (for clarity, the facets of the D study are placed in parentheses after the $\hat{\varepsilon}_\rho^2$, with an asterisk denoting a fixed facet). $\hat{\varepsilon}_\rho^2(X)$ has been defined for this example

by equation 1:

1. Persons random, items random:

$$\hat{\varepsilon}_\rho^2(P,I) = \hat{\sigma}^2(c)/\hat{\varepsilon}_\sigma^2(X) \tag{7}$$

2. Persons random, items fixed:

$$\hat{\varepsilon}_\rho^2(P,I^*) = [\hat{\sigma}^2(c) + \hat{\sigma}^2(ci)/n_i']/\hat{\varepsilon}_\sigma^2(X) \tag{8}$$

3. Persons fixed, items random:

$$\hat{\varepsilon}_\rho^2(P^*,I) = [\hat{\sigma}^2(c) + \hat{\sigma}^2(p:c)/n_p']/\hat{\varepsilon}_\sigma^2(X) \tag{9}$$

For this particular design, $\hat{\varepsilon}_\rho^2(P,I^*)$ is essentially equivalent to what is classically referred to as inter-rater reliability (Ebel, 1951). Similarly, $\hat{\varepsilon}_\rho^2(P^*,I)$ is essentially equivalent to inter-item reliability (Cronbach, 1951). There is no classical counterpart to $\hat{\varepsilon}_\rho^2(P,I)$.

The Generalizability of student ratings of instruction: An application

A major tool for evaluation is the rating scale, where typically a rater evaluates the target of the evaluation in terms of the content of a series of items. Also, as mentioned previously, often it is not persons per se who are being evaluated but collections of persons or other entities (e.g., classes, schools, school districts). Student ratings of instruction is an evaluation technique which incorporates both the use of rating scale information and an evaluative target other than individual persons. As an example application, generalizability research in this area will be presented here (see Gillmore, et al, 1976, 1978 and Kane, et al, 1976).

Consider a typical design under which college-level ratings by students are collected. Commonly, there exists an instrument which is used campus-wide or department-wide. Students enrolled in a given class rate that class using

the items contained in the instrument. For all practical purposes, a different set of students rate each class. Thus, an easily utilized G study design is students or persons nested within classes and crossed with items. This design can be denoted (p:c) x i and is exactly the design which has been used as an example throughout the Formulation section.

Normally, a G study would be designed and run especially for estimation of the various variance components. In the present application this is typically not a practical approach. For one very good reason, investigators do not have control over the users of the instrument in most studies. One can neither assign students to classes nor compel instructors to use the instrument. For another reason, there can be the luxury of an existent accumulation of data which has been collected under this design.

In the specific G study to be presented here for illustration, a random sample of 42 classes were drawn from all of the classes in which the University of Washington Instructional Assessment System (IAS) (Gillmore, 1974) had been used during the academic year, 1974-75, subject to the restriction that each class had to be rated by at least 14 students. For those classes within the sample with more than 14 students, a random sample of 14 was chosen for subsequent analysis. Finally, the IAS contains 5 distinct forms, but 11 items are common to all forms. For the G study, attention was restricted to these 11 items. Thus, through random sampling of the existing data, a completely balanced design was obtained, with 42 classes, 14 students per class, and 11 items. This selection was carried out in order to achieve a balanced design, because, unfortunately, procedures for deriving variance component estimates from unbalanced designs are complex and not well understood.

The results of the G study, then, are reproduced in Table 2. These results are quite similar to those reported in Kane, et al (1976), for two different

instruments used at a different university. Inspection of this table reveals relatively large variance component estimates for both the persons within classes effect and the persons by items within classes effect. Small variance component estimates are found for the items and items by classes effect. Finally, the classes effect has an associated variance component estimate of moderate magnitude.

Table 2 about here

As was stated earlier, in defining a D study the decision-maker must specify the object of measurement, the universe of generalization, the number of conditions to be sampled for each facet, and the experimental design. Let us examine each of these in turn. The object of measurement in this context will almost surely be the class, since the express reason for the students to be responding to the items is to evaluate the class.

With the class as the object of measurement, it follows directly that the facets of the study are persons and items. Each of these can be either considered as fixed or random in defining the universe of generalization, and how each is considered is, of course, dependent upon the purposes of the D study. For the items facet, if the measurement is being taken to provide evidence of some global quality, then the particular items used are not of direct interest. Rather, we would wish to generalize our results to all possible items which could have been used, i.e., to the entire item domain. It is only when the particular items selected completely specify the information sought that it makes sense to treat items as fixed. In the present context, this might be the case for items whose purpose is to be finely diagnostic. A small set of items might completely specify a teaching behavior and, thus, present no need to generalize beyond the set.

It is difficult to present a satisfying argument for treating the persons facet as fixed within the normal context of college level instruction. The average ratings which a particular set of students give to a class is not typically of primary interest. Rather, what is of interest is the average ratings of the class by all students who might have taken it. Decisions about classes would seldom be based on the responses of one particular set of students since, were the class to be offered again, it would be to an entirely or, hopefully, almost entirely new set of students. If results could not be generalized beyond the specific students enrolled in the given class offering, the value of the data for decision making would be highly questionable.

Another decision to be made in defining a D study is the number of conditions of each facet to be sampled. For the items facet, the number of conditions would probably be dependent on the given instrument. However, if one is not locked into using a given instrument or if the G study results can be used in the instrument development phase, then by forecasting generalizability coefficients for differing numbers of items, a decision about adequate instrument length can be made depending upon desired dependability levels. The number of conditions for the persons facet, on the other hand, is outside of the control of the decision-maker. One does not usually determine class size on the basis of what is necessary for dependable evaluation but for pedagogical and practical reasons (e.g., how many students happen to enroll in the class). However, once the G study has been completed, a D study generalizability coefficient estimate is available for any number of students and items. In actuality, two classes rated by differing numbers of students receive results of differing dependability.

The final decision, the experimental design, was essentially made at the front end in this example. That is, a G study design was chosen for

practical reasons which is the same as that under which decisions are made. Hence, the design for the D study was (p:c) x i.

Decisions involving rating scales are almost always made in a comparative mode, and attention here will be restricted only to this kind of interpretation. Decisions based on absolute scores would certainly be possible for scales having meaningful response positions, and that which will be presented subsequently in the section on testing with regard to domain-referenced and mastery testing could be applied here as well.

The general formula for the expected observed score variance for D studies using the (p:c) x i design, with classes as the object of measurement, was given in equation 1. How the expected observed score variance is partitioned into universe score variance and error variance for comparative judgments depends upon definition of the universe of generalization. Earlier, it was argued that usually the persons and the items facet should be considered as random. However, decision-making situations could exist which would suggest one or the other of the facets be considered fixed. Eliminating the trivial case where both facets are treated as fixed, there are three possible generalizability coefficients depending upon whether both facets are considered random or only one is.

Table 3 contains formulae for the universe score variance and the error variance for making comparative judgments for each of these cases. As before, to add clarity the facets of the design are included in capital letters in parentheses after $\hat{\varepsilon}^2$, and facets treated as fixed are further denoted with an asterisk.

Table 3 about here

It should be clear by now that a unique generalizability coefficient estimate exists for each possible combination of the number of levels of

persons and of items. Estimates for each of these coefficients for selected numbers of conditions of the facets are presented in Table 4. The interested reader can determine the estimates which result from any number of conditions for either students or items, casting oneself into the role of the decision-maker.

Table 4 about here

One can notice in viewing the entries of Table 4 that $\hat{\mathcal{E}}_\rho^2(P,I)$ is always smaller than either $\hat{\mathcal{E}}_\rho^2(P,I^*)$ or $\hat{\mathcal{E}}_\rho^2(P^*,I)$, for equivalent numbers of persons and items. This follows directly from the fact that all three coefficients have the same denominator, $\hat{\mathcal{E}}_\sigma^2(X)$, and the latter two have an additional term in the numerator. $\hat{\mathcal{E}}_\rho^2(P,I^*)$ does not seriously over-estimate $\hat{\mathcal{E}}_\rho^2(P,I)$ for equivalent numbers of persons and items except where the number of persons is large and the number of items is very small. This is a direct result of the relative small variance component estimate associated with the class by item interaction. Finally, $\hat{\mathcal{E}}_\rho^2(P,I^*)$ is considerably larger than either of the other two coefficients for comparable numbers of persons and items, except where the number of persons is large. This is because of the relatively large variance component estimate for the persons within classes effect.

What are some of the conclusions which might be reached based on the various dependability estimates of differing D studies? If one were to be cast in the role of an administrator, decisions might have to be made about the promotion of teachers or the continuation of courses.[4] In such a case, the universe of generalization would probably be appropriately defined as an essentially infinite set of items and students. For such decisions, the importance of having a large number of student raters for achieving dependable results is clear in Table 4. Small classes do not yield particularly dependable results. For a small class, say less than 15 or 20, a verdict might await the

ratings of subsequent offerings. One can also readily see that the instrument does not have to be long, five to ten items seems to suffice, and that increasing the number of items beyond this number does not compensate for too few students.

If the decision-maker were the teacher being rated, the purpose of the evaluation might very well be instructional improvement. On items covering very specific behavior, the teacher might feel confident in the meaningfulness of restricting the universe of generalization to the specific items on the instrument. In so doing, one item appears to yield dependable results if a sufficient number of students rate the instructor on that item.

These are two of the many possible perspectives from which the above generalizability analysis could be viewed. How specific D studies are, in fact, conducted and interpreted depends upon the purpose of the evaluation. Again, generalizability theory requires the decision-maker to specify that purpose.

Confidence Intervals

Thus far, the D study statistics discussed have related to an evaluation of a measuring instrument's general properties. Within practical evaluation contexts, a decision-maker would certainly desire an instrument with high demonstrated dependability for the decisions to be made and, inversely, with relatively little error variance. However, one might wish to go beyond these general properties of the instrument to questions about individual units of the object of measurement, such as a person or a class. It is at this level, after all, that decisions are made. For example, in comparing two classes, one might wish to know the likelihood that the universe score of Class A is larger than that of Class B, given some discrepancy in their respective observed

scores. Or, one might wish to know how likely a person's universe score is greater than a cutoff score, given some observed score. Considerations such as these involve us in discussions of confidence intervals.

The square root of the error variance for making comparative judgments, $\sigma(\delta)$, is analogous to what in classical test theory is known as the standard error of measurement. This index of error is appropriate to use in constructing confidence intervals around an observed mean score when it is the relative position of that score which is of interest. If the value of the observed mean score itself is of interest, the square root of the average error within the object of measurement, $\sigma(\Delta)$, is the appropriate error index to use in constructing intervals.

In classical test theory, confidence intervals are constructed by multiplying the standard error of measurement by a constant from the t tables, depending upon what level of confidence is sought, and the product is then added to the observed mean score to obtain the higher value of the range and subtracted from the observed mean score to obtain the lower value of the range. This procedure is followed at all points on the score scale, and yields confidence intervals which are symmetric and of uniform width. Cronbach, et al (1972, p. 127ff) argue convincingly that the straight-forward application of this procedure when individuals are the object of measurement is unforgivably inaccurate for scores which depart markedly from the group mean. Use of the information which can be gained from prior knowledge or intelligent guesses about the distribution of the population of scores from which the given score came along with the error index can give a much more accurate and useful interval and one which will not be symmetric around the observed score mean unless that mean is equal to the population mean.

When the object of measurement represents a mean score over persons, e.g., classes, confidence intervals constructed to be of uniform width and symmetric around the class mean in the manner described above, either as a function of $\sigma(\delta)$ or $\sigma(\Delta)$, are generally much more accurate along various points on the score scale than those similarly computed for individuals. Classically, the error standard deviation used for confidence intervals around class means is the standard error of the mean, which is strictly based on a persons within classes variance estimate. Using $\sigma(\delta)$ or $\sigma(\Delta)$ rather than the standard error of the mean allows variance estimates from additional sources to enter into the error variance when appropriate.

## The Testing of Persons

Along with rating scales, the testing of persons represents a major source of data within evaluation research. Occasionally achievement tests are used to evaluate teachers, programs, and even the success of school desegregation (Bradley and Bradley, 1977). In the main, however, tests are intended to evaluate the persons to whom they are administered, and it is this function to which this section is addressed. However, discussion could be easily extended to decision-making contexts in which an object of measurement other than persons is appropriate.

In this presentation of the application of generalizability theory to testing, the author is indebted to the work of Robert Brennan and Michael Kane,

presented in Brennan and Kane (1977a, and 1977b) and Kane and Brennan (1977), and synthesized, illustrated, and extended in Brennan (1978).

In the current literature, distinctions are drawn between norm-referenced and domain or criterion-referenced tests (see Hambleton, Swaminathan, Algina, and Coulson, 1978). Reasonably clear differences exist in the purposes which these two types of tests serve and how each is constructed. Norm-referenced tests have as a goal discrimination among testees and, hence, are developed to maximize the variance of the test scores. The items of norm-referenced tests tend to be relatively general in nature. In contrast, domain-referenced tests are designed to assess specific competencies. Assessment of individual differences is of minor importance. Items are chosen to be specifically tied to desired behaviors or skills. Following Brennan and Kane (1977a), a domain-referenced test with a single absolute cutting score which defines successful performance is termed a mastery test. Licensing examinations are one class of tests having this characteristic; minimum competency examinations are another.

It is important to recognize that while these types of tests differ in characteristics, a more fundamental differentiation is in terms of the interpretation of their results. There are at least two distinct ways in which test scores can be interpreted. First, a person's score, averaged over the set of items of a given test, can be interpreted relative to the average scores of other persons either taking the test simultaneously, or having taken it previously as a norm group. A test, so interpreted, is a norm-referenced test. Second, a person's score can be interpreted with reference back to the item domain itself regardless of how others have performed on the same test. The standards in this case are relative to the item domain and tests given this interpretation are known as "domain-referenced." A mastery test, then, is a subclass of domain-referenced tests in which the interpretation is relative

to a single cutoff score, and that cutoff score is chosen with regard to the item domain and not with regard to the relative performance of a norm group.

Confusions arise over the fact that results from the same test can be given either a domain-referenced or a norm-referenced interpretation, although a single test would probably not adequately fulfill both functions. Thus, strictly speaking, it is not completely proper to speak of norm-referenced and domain-referenced tests. However, the main-stream literature in the area appears content to do so and that lead is followed here. The reader should keep in mind, though, that the fundamental distinction is made on the basis of how the test is to be interpreted.

In viewing the approach of generalizability theory to these three types of tests, the simple persons crossed with items (p x i) design will be used for illustration. More complex designs could have been chosen, such as persons crossed with items, items nested within subtests, subtests nested within tests, etc., and the following development can be easily extended to these more complex designs (see Brennan, 1978). The G study conducted using the p x i design yields three variance component estimates: $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(pi)$. The object of measurement will be assumed to be persons, as would be the case for almost any application using this design. Items is the only facet of the study and will be considered as random throughout this discussion.

When treating the items of a test as a random facet within generalizability theory, the items used in the G study are assumed to be randomly sampled from some well-defined universe or domain of items, and the items of the D study are assumed to be randomly sampled from the same domain. Clearly this item domain must be meaningful if generalization to it is to have any meaning. If the domain is a hodge-podge of items, the average over which does not measure any unitary ability, then assessment of the dependability of the test has

little value. An index of dependability is meaningful only insofar as the universe of generalization, in this case the item domain, is meaningful. This is true whether scores are to be interpreted relative to the item domain or relative to other scores, and, for the former, whether there is a single cutoff score of interest or not.

More will be said on the universe of generalization for tests in the "Concluding Comments" section to follow. We now turn to specific discussion of the three types of tests from the perspective of generalizability theory. For the basic (p x i) design, the results of this analysis for norm referenced tests are fundamentally the same as those of classical test theory. This is not true for domain-referenced and mastery testing.[5]

Norm-referenced tests. What is of interest in norm-referenced tests is not the person's absolute score, but the person's score relative to the score of others. The universe score variance is estimated by the variance component estimate for the persons main effect alone. Importantly, the proper error variance to use in this context is that for making comparative judgments, $\sigma^2(\delta)$. For the present D study design, only the variance component associated with the persons by items interaction comprises this statistic. The proper index of dependability is the generalizability coefficient which is defined as the universe score variance divided by the expected observed score variance:

Insert Equation 10 here

For the p x i design with persons as the object of measurement and the items facet treated as random, $\widehat{\mathcal{E}\rho}^2(I)$ is equivalent to KR-20 (Kuder and Richardson, 1937) for dichotomously scored items and Cronbach's alpha (Cronbach, 1951) in the more general case. Furthermore, the magnitude of $\widehat{\mathcal{E}\rho}^2(I)$ is altered as a function of the number of D study items upon which decisions will be based exactly as is the case through application of the Spearman-Brown prophesy

formula. In general, the Spearman-Brown formula holds only in situations in which there is no more than one random facet.

It is especially important to note that the variance component estimate associated with the item main effect is not a source of error variance for making comparative decisions. These decisions are based on the score of each person, averaged over all items. Thus, differences in item difficulty make no difference in the deviation score for each person.

Domain-referenced tests. Attention will be restricted here to the general case in which there is no particular cutoff score of interest, but interpretations are related back to the item domain rather than being related to the performance of others. Mastery interpretation will be taken up subsequently.

The universe score variance is defined in the same way for domain-referenced tests as it is for norm-referenced tests. For the p x i design, the universe score is estimated by the variance component associated with the main effect of persons as above. However, the proper error variance for this interpretation is the error variance within the object of measurement, $\sigma^2(\Delta)$ (see Brennan and Kane, 1977a). The error variance within the object of measurement differs from the expected error variance for making comparative judgments in that the former includes all variance component estimates which contain a random facet index, while the latter includes only those variance component estimates containing both a random facet index and the object of measurement index. For the p x i design, the error variance within the object of measurement is estimated as follows:

Insert Equation 11 here

For this design, the difference between $\hat{\sigma}^2(\delta)$ and $\hat{\sigma}^2(\Delta)$ is the inclusion of the term, $\hat{\sigma}^2(i)/n_i'$ in the latter. Intuitively, this is because the

judgments about persons' abilities are based on their scores on a set of items representing a particular random sample from the item domain. These scores would be expected to change as different sets of items from the same domain are randomly chosen. Under a norm-referenced interpretation, these changes are unimportant because persons' relative positions are unchanged. But under a domain-referenced interpretation, these changes can be important. If all items were equally difficult, the particular items chosen would not influence persons' absolute scores, $\hat{\sigma}^2(i)$ would equal zero, and $\hat{\sigma}^2(\delta)$ would equal $\hat{\sigma}^2(\Delta)$. Otherwise, $\hat{\sigma}^2(\Delta)$ will be larger than $\hat{\sigma}^2(\delta)$ because the specific set of items chosen influence persons' absolute scores and, hence, there is added error variance from this source.

The index of dependability for domain-referenced tests, then, is the ratio of universe score variance to the sum of universe score variance and error variance within the object of measurement. The denominator for this index is generally not equal to the expected observed score variance and, hence, it is not a generalizability coefficient as defined above. Following Brennan (1978), the generic label "index of dependability" will be maintained here and denoted by $\Phi$. In the p x i design which has been used above, the index of dependability for domain-referenced tests is estimated as follows:

Insert Equation 12 here

Mastery tests. To reiterate, a test with a mastery interpretation is defined as a domain-referenced test with a single cutting score. Persons who obtain or exceed this score are said to have mastered the content represented by the item domain. Persons scoring below the mastery level are said to have failed to master the content.

The proper error term for the index of dependability remains the error variance within the object of measurement, as in the general case for

domain-referenced interpretations. However, interest is now focused on the absolute performance of a person in relation to a cutting score, which will be denoted $\lambda$. As a result, a term, first introduced by Livingstone (1972), must be added to both the numerator and the denominator of the formula for the dependability index. This term reflects the difference between the mean of the test and the cutting score. If the population mean is denoted $\mu$, and the cutoff score $\lambda$, the formula for the dependability of mastery tests is as follows (Brennan, 1978):

Insert Equation 13 here

The formula for estimating $\Phi(\lambda)$ is somewhat more complex because simply substituting the sample mean for $\mu$ leads to a biased estimate of the squared difference between the population mean and the cutoff score (see Brennan and Kane, 1977).

If $\lambda = \mu$, the index of dependability for mastery tests is equivalent to that for domain-referenced tests with no particular cutoff score. Otherwise, as $\lambda$ departs from $\mu$, the dependability of the test for making mastery decisions improves. The most difficult situation for mastery decisions is when examinees cluster around the cutoff score. The easiest is when examinees cluster elsewhere.

## Concluding Comments

If there is a single concept which provides the heart of generalizability theory, it is the idea of a universe of admissible observations and the closely related universe of generalization. To review, Cronbach, et al (1972) used the phrase "universe of admissible observations" with reference to G studies and "universe of generalization" with reference to D studies. The universe of admissible observations refers to the factors which one chooses

to systematically take into consideration within the G study and the range of conditions which may be sampled from each. A potential facet of a D study can be treated in one of three ways in the G study. First, it can be systematically represented by sampling two or more of its admissible conditions, and thus be a factor in the G study. Secondly, it can be held constant within the G study, i.e., only one condition of the potential facet is admissible for all observations. Finally, it can be allowed to vary without any controls and thus become an undifferentiated component of random error. Consider, for example, occasions as a potential facet. For the G study, two or more occasions could be systematically sampled in a crossed or nested design. Alternatively, all measurements could be taken at the same point in time. Finally, measurements could be taken over a reasonable length of time, say three weeks, with no direct attention to the occasion of the measurement. Each of these three possibilities presents a different universe of admissible observations, and the latter two a priori limit the universe of generalization.

In the one application presented above, i.e., student instructional ratings, discussion was somewhat remiss in not carefully specifying the characteristics of the universes of generalization, other than in fairly general terms. This was a conscious choice so as to avoid getting bogged down needlessly in what was intended to be merely an example. However, one's conclusions are clearly very intimately tied to the specification of the universe of generalization. As Cronbach, et al (1972) stated, "Investigators often choose procedures for evaluating the reliability that implicitly define a universe narrower than their substantive theory calls for. When they do so, they underestimate the 'error' of measurement, that is, the error of generalization" (p. 352).

The assumptions of generalizability theory are considerably weaker than those of classical test theory, and, in fact, come down to one central assumption: that conditions are randomly sampled when data are collected (Cronbach, et al 1972, p. 357). While this weakening introduces flexibility and potentially greater understanding into the analyses of one's measuring device, it also lays bare the importance of the characteristics of the sample. In practice, true random samples are seldom achievable because the universe is typically innumerable--all admissible observations cannot be listed. It would be rare, for example, to be able to completely list all possible test items in a given content domain, even if the characteristics of those test items were narrowly specified.

This problem can be reduced somewhat by viewing samples from universes as operationally defined. Cronbach, et al (1972) described operationally defined samples as follows: "The essential requirement for reasonable application of sampling theory is that the universe be defined clearly enough that one can recognize elements belonging to it" (p. 368). Thus, according to this view, all possibilities do not have to be listed with random samples then drawn from the list. Rather, standards can be articulated which allow any investigator to discriminate those elements (e.g., persons or items) which are admissible from those which are not.

The aspect of the problem which this perspective leaves unresolved is that the persons used in a study are often those most available, the items are those determined to be best by expert judgment, etc. For example, an instrument to be used in a national evaluative study might be pilot tested in a single location. A test might consist of the best items from prior item analyses. In fact, the universe often becomes defined as containing a population of persons like those in the sample, a universe of items like

those on the test, etc. Yielding to loosely defined universes first of admissible observations and consequently of generalization stands in the way of effective theory development (e.g., determination of which facets make an important difference), and often weakens the essential G study D study link. This is not a problem inherent in the formulation of generalizability theory or unique to it. It rather suggests a potential misuse or under-use of the power of the theory.

Even when the universe of admissible observations is clearly operationalized, the exact nature of the universe of generalization may be vague. For example, a clear rule for determination of whether a given item is an admissible condition for the item facet may exist, but the universe from which the particular set of items of the G study can be considered a random sample may still be incompletely understood. In discussing this reality, Cronbach, et al (1972) appealed to the bridge analogy of Cornfield and Tukey (1956), not as a solution to the problem but as a perspective on it. In this paper, Cornfield and Tukey suggested that the leap from actual observations to conclusions has two components, one of which is statistical and one of which is subject-matter related. "If we take the simile of the bridge crossing a river by way of an island, there is a statistical span from the near bank to the island, and a subject-matter span from the island to the far bank... It may often be worthwhile to move the island nearer to the distant bank, at the cost of weakening the statistical span--particularly when the subject-matter span is weak" (pp. 912-913). In the present context, their point might well be in support of the superiority of generalizing to a vague universe considered to be essentially infinite over focusing attention on a universe which is considered to be fixed and, thus, well-defined. The exact location of the island under the former strategy may be uncharted, but

at least we can have confidence that it is near the subject-matter shore which is where our greatest interests lie.

How one defines the universe of generalization can have direct implications for either content or construct validity. In a testing application, specification of the universe of admissible observations for the items factor can be cast as a classical content validity problem. When a given test is analyzed for content validity, items are seldom considered only in terms of whether or not each should be present in an undifferentiated specification of acceptable test content. In addition, the acceptable content of a test is usually divided into more content specific sub-areas or strata. A test, then, is evaluated with regard to how thoroughly its items cover the various strata. In this regard, items may be better considered as sampled within strata than as sampled from a universe of more diverse content. "A proposal to sample items from a broad domain at random is generally but not always a sign that one's understanding is crude" (Cronbach, et al 1972).

This more sophisticated view of the universe of generalization of a test leads to extension of the basic design in which persons are crossed with items to a design in which items are nested within strata as well as being crossed with persons. This design could be denoted as follows: p x (i:s). In the universe of generalization, the strata facet would be most often viewed as fixed, since usually the strata completely specify the content area and there is no intention to generalize beyond them. On the other hand, one would expect an intent to generalize the item facet beyond those items contained within the strata and, thus, the items facet should be considered as random. Importantly, generalizability theory is presented with no difficulty in treating the above design. In fact, further hierarchical structuring is possible, e.g., substrata within strata.

Resulting indices of dependability are not, of course, indices of content validity, and the dependability of a D study with persons as the object of measurement still relates to the test as a whole. However, the relative magnitudes of the variance component estimates may be of direct interest. For example, a relatively large strata by persons variance component estimate might suggest that the specification of the item domain needs to be rethought because abilities which are not unitary in nature are being measured.

The universe of generalization can also have implications for construct validity. Earlier in this section, loosely specified universes were contrasted with more carefully defined ones. Universes can also be distinguished in terms of how narrowly or broadly they are defined. When a universe is narrowly defined, consisting of all highly similar measures, e.g., all multiple choice test items, the resulting dependability estimate is clearly what is classically known as a reliability estimate. But, when the universe contains diverse measures, preferably with the type of measurement being a facet of the D study, the resulting dependability estimate begins looking like an index of construct validity. The point is simply stated by Campbell and Tyler (1957) as quoted in Cronbach, et al (1972). "Reliability is epitomized by the correlation of two specifications of a construct through maximally similar approaches. Construct validity is epitomized by the correlation between two or more specifications of a construct maximally different in apparatus or method" (p. 91). This is particularly relevant to evaluation research in which multiple indicators of a program's quality are obtained. As these indicators become more diverse, validity considerations become more prominent.

This should not lead one into thinking that a dependability estimate describes either reliability or construct validity and that the two are easily discriminated. The distinction between the two is often quite fuzzy. Funda-

mentally, the question goes right back to the universe of generalization and how the decision-maker views it. The extent to which one is satisfied that the universe of generalization is, in fact, what he or she wishes to know about the object of measurement, as opposed to being a partial indicator thereof, is the extent to which construct validity is being studied.

Cronbach, et al (1972) made the following significant observation which neatly sums up the present discussion:

> The operational definition of an indicator or a body of content to be sampled defines a universe. Generalizability analysis indicates how closely different realizations of the definition correspond to each other and to the universe score. The more definitively the developer of a measuring procedure describes the universe he has in mind, the greater degree of generalizability he can hope to attain with a sample of practicable size. The question remains whether he has defined a universe particularly worthwhile investigating. To answer this, one asks for convergence of diverse indicators. One does not move far toward construct validation until he has established convergence of indicators of different kinds, each representing its own universe of admissible operations (p. 381).

## FOOTNOTES

[1]The author wishes to acknowledge the important contributions to this paper made by Drs. Robert L. Brennan, Michael T. Kane, and Daniel W. King through their careful and insightful critiques of a preliminary version.

[2]Some might suggest that even this is not what the teacher wants, because we have limited ourselves to a certain type of test. In this view, what is really sought is a quantification of the ability of interest which transcends any particular measurement procedure. This concern propels us immediately into considerations which are classically labeled validity. Discussion of this topic is deferred to the concluding section of this paper.

[3]It might be useful to try to anticipate a possible source of confusion which individuals often suffer upon first exposure to the distinction between fixed and random facets within generalizability theory. In the analysis of variance literature, an often cited rule of thumb for determining whether a given factor should be treated as random or fixed is to consider what levels of that factor would be chosen if the experiment were to be replicated. This rule states that the factor should be considered fixed if the same levels would be chosen and random if different levels would be chosen. The rule is useful for the determination of fixed vs. random factors in analyses of variance, but it is not relevant to the equivalent decision in defining D studies. A G study might be replicated using the same items, but D study generalization applied to an essentially infinite item domain would still be appropriate. Also, a G study might be replicated using a different set of items from the same domain, but the decision-maker might still choose not to generalize the D study results beyond the specific items used in the D study.

[4]Strictly speaking, decisions about teachers require that teachers be the object of measurement. Similarly for courses. Classes as the object of measurement confounds these two. For a discussion of this and a report of generalizability studies with teachers and courses as objects of measurement, see Gillmore, Kane, and Naccarato (1978).

[5]The discussion of the generalizability of domain-referenced test results to be presented here is based on a definition of error as differences between observed and universe absolute scores. An alternative approach is the view of errors as misclassifications and the use of Cohen's $\kappa$ (Cohen, 1960) or some variant as an index of reliability (see Hambleton, et al, 1977).

Table 1

Random Effects

G Study and D Study Variance Components for the (p:c) x i Design with Classes As the Object of Measurement.

| Effects | G Study Variance Component | D Study Variance Component |
|---|---|---|
| c | $\sigma^2(c)$ | $\sigma^2(c)$ |
| p:c | $\sigma^2(p:c)$ | $\sigma^2(p:c)/n_p'$ |
| i | $\sigma^2(i)$ | $\sigma^2(i)/n_i'$ |
| ci | $\sigma^2(ci)$ | $\sigma^2(ci)/n_i'$ |
| pi:c | $\sigma^2(pi:c)$ | $\sigma^2(pi:c)/(n_p'n_i')$ |

Table 2

The G Study Analysis of Variance Results (from Gillmore, et al 1976).

| Effect | SS | DF | MS | $\hat{\sigma}^2$ |
|---|---|---|---|---|
| Classes (c) | 1039.29 | 41 | 25.35 | .13 |
| Persons within classes (p:c) | 2825.64 | 546 | 5.18 | .43 |
| Items (i) | 74.65 | 10 | 7.46 | .01 |
| ci | 455.07 | 410 | 1.11 | .05 |
| pi:c | 2505.00 | 5460 | .46 | .46 |

Table 3

Universe Score and Error Variance Formulae for Three Generalizability Coefficients (p:c x i design).

| Facet | | | | |
|---|---|---|---|---|
| Persons | Items | Generalizability Coefficient | Universe Score Variance - $\hat{\sigma}^2(\tau)$ | Error Variance $\hat{\sigma}^2(\delta)$ |
| Random | Random | $\widehat{\mathcal{E}\rho}^2(P,I)$ | $\hat{\sigma}^2(c)$ | $\hat{\sigma}^2(ci)/n_i' + \hat{\sigma}^2(p:c)/n_p' + \hat{\sigma}^2(pi:c)/n_p'n_i'$ |
| Random | Fixed | $\widehat{\mathcal{E}\rho}^2(P,I^*)$ | $\hat{\sigma}^2(c) + \hat{\sigma}^2(ci)/n_i'$ | $\hat{\sigma}^2(p:c)/n_p' + \hat{\sigma}^2(pi:c)/n_p'n_i'$ |
| Fixed | Random | $\widehat{\mathcal{E}\rho}^2(P^*,I)$ | $\hat{\sigma}^2(c) + \hat{\sigma}^2(p:c)/n_p'$ | $\hat{\sigma}^2(ci)/n_i' + \hat{\sigma}^2(pi:c)/n_p'n_i'$ |

Table 4

Generalizability Coefficient Estimates for Various D Studies

| | | $\hat{\mathcal{E}}_\rho^2(P,I)$ | $\hat{\mathcal{E}}_\rho^2(P,I^*)$ | $\hat{\mathcal{E}}_\rho^2(P^*,I)$ |
|---|---|---|---|---|
| $n_p' = 1$ | $n_i' = 1$ | .12 | .17 | .52 |
| $n_p' = 5$ | $n_i' = 5$ | .53 | .57 | .88 |
| $n_p' = 10$ | $n_i' = 5$ | .68 | .73 | .90 |
| $n_p' = 10$ | $n_i' = 10$ | .71 | .74 | .95 |
| $n_p' = 25$ | $n_i' = 1$ | .60 | .83 | .68 |
| $n_p' = 25$ | $n_i' = 5$ | .81 | .87 | .91 |
| $n_p' = 25$ | $n_i' = 10$ | .84 | .88 | .96 |
| $n_p' = 25$ | $n_i' = 25$ | .87 | .88 | .98 |
| $n_p' = 100$ | $n_i' = 25$ | .95 | .97 | .98 |

## EQUATIONS

(1) $\hat{\mathcal{E}}_{\sigma}^2(X) = \hat{\sigma}^2(c) + \hat{\sigma}^2(p:c)/n_p' + \hat{\sigma}^2(ci)/n_i' + \hat{\sigma}^2(pi:c)/n_p'n_i'$

(2) $\hat{\sigma}^2(\delta) = \hat{\sigma}^2(p:c)/n_p' + \hat{\sigma}^2(ci)/n_i' + \hat{\sigma}^2(pi:c)/n_p'n_i'$

(3) $\hat{\sigma}^2(\delta) = \hat{\sigma}^2(ci)/n_i' + \hat{\sigma}^2(pi:c)/n_p'n_i'$

(4) $\hat{\sigma}^2(\varepsilon) = \hat{\sigma}^2(p:c)/n_p' + \hat{\sigma}^2(pi:c)/n_p'n_i'$

(5) $\hat{\sigma}^2(\Lambda) = \hat{\sigma}^2(p:c)/n_p' + \hat{\sigma}^2(i)/n_i' + \hat{\sigma}^2(ci)/n_i' + \hat{\sigma}^2(pi:c)/n_p'n_i'$

(6) $\hat{\sigma}^2(\tau) = [\hat{\sigma}^2(\tau)/\hat{\mathcal{E}}_{\sigma}^2(X)]\,[\hat{\mathcal{E}}_{\sigma}^2(X) - \hat{\sigma}^2(\tau)]$

(7) In body of report

(8) In body of report

(9) In body of report

(10) $\hat{\mathcal{E}}_{\rho}^2(I) = \hat{\sigma}^2(p)/[\hat{\sigma}^2(p) + \hat{\sigma}^2(pi)/n_i']$

(11) $\hat{\sigma}^2(\Lambda) = \hat{\sigma}^2(i)/n_i' + \hat{\sigma}^2(pi)/n_i'$

(12) $\hat{\Phi} = \hat{\sigma}^2(p)/[\hat{\sigma}^2(p) + \hat{\sigma}^2(i)/n_i' + \hat{\sigma}^2(pi)/n_i']$

(13) $\Phi(\lambda) = [\sigma^2(p) + (\mu - \lambda)^2]/[\sigma^2(p) + \sigma^2(i)/n_i' + \sigma^2(pi)/n_i' + (\mu - \lambda)^2]$

## REFERENCES

American College Testing Program. Assessing Students on the Way to College, Volume 1, Iowa City: ACT Publications, 1973.

Bradley, L.A. and Bradley, G.W. The academic achievement of black students in desegregated schools: a critical review. Review of Educational Research, 1977, 47, 399-449.

Brennan, R.L. Generalizability analyses: principles and procedures. ACT Technical Bulletin, No. 26, Iowa City: The American College Testing Program, 1977.

Brennan, R.L. Extensions of generalizability theory to domain-referenced testing. ACT Technical Bulletin, No. 30, Iowa City: The American College Testing Program, 1978.

Brennan, R.L. and Kane, M.T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289(a).

Brennan, R.L. and Kane, M.T. Signal/noise ratios for domain-referenced tests. Psychometrika, 1977, 42, 609-625(b).

Campbell, D.T. and Tyler, B.B. The construct validity of work-group morale measures. Journal of Applied Psychology, 1957, 41, 91-92.

Cardinet, J., Tourneur, Y., and Allal, L. The symmetry of generalizability theory: applications to educational measurement. Journal of Educational Measurement, 1976, 13, 119-135.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cornfield, J. and Tukey, J.W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of generalizability for scores and profiles. New York: John Wiley and Sons, 1972.

Cronbach, L.J., Rajaratnam, N., and Gleser, G.C. Theory of generalizability: a liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Ebel, R.L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Edwards, W., Guttentag, M., and Snapper, K. A decision-theoretic approach to evaluation research. In Struening, E.L. and Guttentag, M. (eds.), Handbook of Evaluation Research, Volume 1, Beverly Hills: Sage Publications, Inc., 1975.

Fisher, R.A. Statistical Methods for Research Workers. London: Oliver and Boyd, 1925.

References (continued)

Fletcher, J. Situation Ethics. Philadelphia: Westminster Press, 1966.

Gillmore, G.M. A brief description of the University of Washington Instructional Assessment System. EAC Reports, Seattle: Educational Assessment Center, University of Washington, 1974.

Gillmore, G.M., Kane, M.T., and Naccarato, R.W. The generalizability of student instructional ratings: general theory and application to the University of Washington Instructional Assessment System. EAC Reports, No. 76-14, Seattle: Educational Assessment Center, University of Washington, 1976.

Gillmore, G.M., Kane, M.T. and Naccarato, R.W. The generalizability of student ratings of instruction: estimation of the teacher and course components. Journal of Educational Measurement, 1978, 15, 1-13.

Gleser, G.C., Cronbach, L.J., and Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. Psychometrika, 1965, 30, 395-418.

Gulliksen, H. Theory of Mental Tests. New York: John Wiley and Sons, 1950.

Hambleton, R.K., Swaminathan, H., Algina, J., and Coulson, D.B. Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, 1978, 48, 1-48.

Kane, M.T., and Brennan, R.L. Agreement coefficients as indices of dependability for domain-referenced tests. ACT Technical Bulletin, No. 28, Iowa City: The American College Testing Program, 1977(a).

Kane, M.T., and Brennan, R.L. The generalizability of class means. Review of Educational Research, 1977, 47, 267-292(b).

Kane, M.T., Gillmore, G.M., and Crooks, T.J. Student evaluations of teaching: the generalizability of class means. Journal of Educational Measurement, 1976, 13, 171-184.

Kennedy, M.M. Findings from the Follow Through planned variation study. Educational Researcher, June 1978, 7, 3-11.

Kuder, G.F. and Richardson, M.W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Lord, F.M. and Novick, M. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

Myers, J.L. Fundamentals of Experimental Design. Boston: Allyn and Bacon, 1966.

Rajaratnam, N. Reliability formulas for independent decision data when reliability data are matched. Psychometrika, 1960, 25, 261-271.

Wallace, I., Wallechinsky, D., and Wallace, A. Book of Lists. West Caldwell, New Jersey: William Morrow and Co., 1977.

Weise, C.H. Evaluation research in the political context. In Struening, E.L. and Guttentag, M. (eds.), Handbook of Evaluation Research, Volume 1, Beverly Hills: Sage Publications, Inc., 1975.

# END

DEPT. OF HEW

NAT'L INSTITUTE OF EDUCATION

ERIC

DATE FILMED

MAR. 13. 1980